Poster I-1

Exploring Life Sciences Data Sources

Lacroix, Zoé[*1], Naumann, Felix[2], Raschid, Louiqa[*3], Vidal, Maria-Esther[4]
[1]Arizona State University, Tempe, AZ, USA; [2]Humboldt University of Berlin, Germany; [3]University of Maryland, College Park, MD, USA; [4]Simon Bolivar University, Venezuela

There has been an explosion of the data that are available to life scientists. While this explosion presents an opportunity, it is accompanied by difficulties in harnessing and exploring this data. Public life science data sources represent a complex link-driven federation of sources. A fundamental problem facing researchers today is correctly identifying a specific instance of a biological entity, e.g., a specific gene or protein, and then obtaining a complete functional characterization of this entity instance by exploring a multiplicity of inter-related sources.

Life science data sources contain data on classes of scientific entities such as genes and sequences. Each source may have data on one or more classes. There is significant diversity in the coverage of these sources. For example, NCBI and EMBL Nucleotide databases have different attributes characterizing (describing) sequences although they both cover the same sequences. Despite being three gene databases, AllGenes, RatMap and the Mouse Genome Database (MGD) cover different datasets. MGD covers mouse genes, when RatMap only contains rat genes. AllGenes contains human and mouse genes, thus overlaps with MGD.

Relationships between scientific objects are often implemented as physical links between data sources. Each physical link between sources may be visualized as a collection of individual links, going from a data object in one source to another data object, in the same or a different source. The physical implementation of these links may vary, e.g., embedded identifiers, URLs, etc. Properties of the relationship such as uni or bi-directional, 1:1 or 1:N, etc. may also vary widely. A scientist is often interested in exploring relationships between scientific objects, e.g., genes and citations. These objects may be retrieved from various data sources, e.g., PubMed for publications. Such an exploration process typically starts from one or more of these available sources, the scientist browsing from one object to the other, following direct links, e.g., a URL, or traversing paths, i.e., concatenations of links via intermediate sources.

Given some start class in source S and target class in source T, there may be multiple alternate paths to navigate from S to T. For instance the query "retrieve the sequences relevant to the citation *Suppression of apoptosis in mammalian cells by NAIP and a related family of IAP gene*s" may be evaluated by navigating on different paths from the start source PubMed to the target source NCBI Nucleotide. A first path consists of the extraction of GenBank identifiers from the MEDLINE format of the PubMed entry (4 GenBank entries are retrieved), while the second path follows the ENTREZ Nucleotide link (9 GenBank entries are retrieved). Each path potentially yields very different results with different properties. This depends on the following: the attributes characterizing each source; the intermediate sources and corresponding entity classes that are traversed in a path; and the contents of each source and each physical link between sources. An example property is the number of data objects of the target class T that are obtained by starting from (a relevant set of) data objects in S. Note that result cardinality may vary based on the choice of the path.

These properties are of interest from a number of perspectives. For example, from a query evaluation viewpoint, one can predict the cost of evaluating a query given some specific sources and paths. This can impact query optimization. One could also choose specific sources and paths depending on some criteria that are evaluated on these properties, e.g., to maximize result cardinality or to maximize the number of attributes. Such criteria impact the domain specific semantics of the results. In this poster, we summarize the approach followed to achieve two research tasks. The first task involves algorithms to explore the search space of links and paths between biological data sources, and to efficiently identify paths that are relevant to a query expressed by a scientist. The second task is to develop a framework to determine the properties that characterize (multiple alternate) links and paths between two sources. Together, these tasks provide a solid foundation to support scientific exploration.